Klasické metody redukce dimenze dat

1. Připomenutí PCA (Principal Component Analysis, rozklad na hlavní komponenty)

2. ICA (Independent Component Analysis, rozklad na nezávislé komponenty)
   [Jutten & Hérault 1991]

3. Fisherova diskriminační analýza

Literatura k tématu

[1] R.O. Duda – P.E. Hart – D.G. Stork. Pattern Classification. John Wiley & Sons, 2001.

[1] A. Hyvärinen – E. Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4-5):411-430, 2000.

PCA: Rozklad na hlavní komponenty

Máme množinu bodů $\mathbf{x}_1, \ldots, \mathbf{x}_k$ v prostoru dimenze d, z bodů vytvoříme $d \times k$ matici $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$

Hledáme nejlepší reprezentaci $\mathbf{X}$ v prostoru dimenze $s < d$

1. Nejlepší reprezentace pro $s = 0$ minimalizuje energii

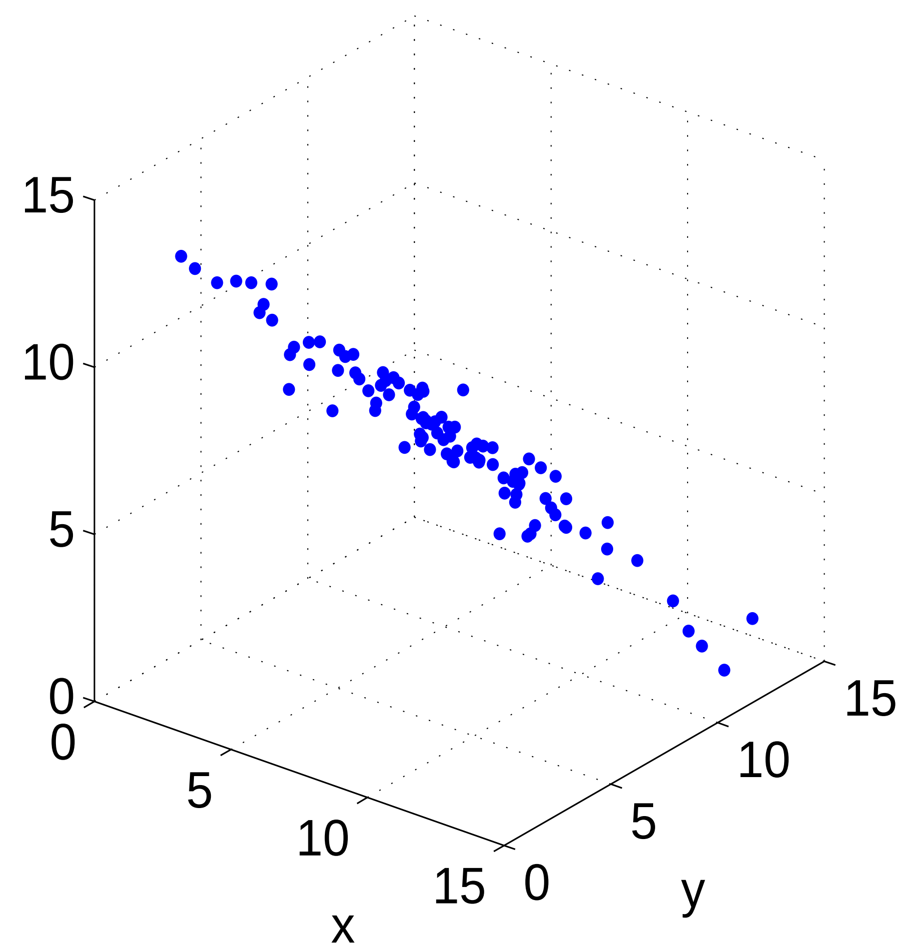
$$J_0 = \sum_{i=1}^{k} \|\mathbf{x}_i - \mathbf{x}_0\|^2,$$

kde řešením je

$$P(\varepsilon \mid x_0) = \prod_{i=1}^{k} p(\varepsilon_i \mid x_0)$$

$$\mathbf{x}_0 = \frac{1}{k} \sum_{i=1}^{k} \mathbf{x}_i \qquad p = \frac{1}{T} e^{-\varepsilon_i^2}$$

důkaz: ⊛1



PCA: pokračování

2. Nejlepší reprezentace při $s = 1$ je průmět na přímku

$$\mathbf{y}_i = \mathbf{x}_0 + t_i \mathbf{w}, \qquad \text{kde } \mathbf{w} \text{ je vektor dimenze } d \text{ a } \|\mathbf{w}\| = 1$$

Parametry t_i, $i = 1, 2, \ldots, k$ a $\mathbf{w}$ minimalizují energii

$$J(t_1, \ldots, t_k, \mathbf{w}) = \sum_{i=1}^{k} \|\mathbf{x}_0 + t_i \mathbf{w} - \mathbf{x}_i\|^2 = \sum_i \|t_i \mathbf{w} - \bar{\mathbf{x}}_i\|^2 = \sum_i (t_i \mathbf{w} - \bar{\mathbf{x}}_i)^\top (t_i \mathbf{w} - \bar{\mathbf{x}}_i) =$$

$$= \sum_i \left(t_i^2 - 2 t_i \, \mathbf{w}^\top \bar{\mathbf{x}}_i + \bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_i \right)$$

z $\frac{\partial J}{\partial t_i} = 0$ dostaneme $t_i = \mathbf{w}^\top \bar{\mathbf{x}}_i = \bar{\mathbf{x}}_i^\top \mathbf{w}$. Po dosazení

$$J(t_1, \ldots, t_k, \mathbf{w}) = \cdots = \sum_i \bar{\mathbf{x}}_i^\top \bar{\mathbf{x}}_i - \mathbf{w}^\top \left(\sum_i \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\top \right) \mathbf{w} = \sigma^2 - \mathbf{w}^\top \mathbf{S} \, \mathbf{w},$$

kde $\mathbf{S}$ je $d \times d$ kovarianční matice $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$

PCA: pokračování

Hledáme minimum J za podmínky $\|\mathbf{w}\|^2 = 1$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\sigma^2 - \mathbf{w}^\top \mathbf{S} \mathbf{w} + \lambda(\mathbf{w}^\top \mathbf{w} - 1)\right)$$

$S w = \lambda w$

z $\frac{\partial J}{\partial \mathbf{w}} = 0$ dostaneme $2\mathbf{S}\mathbf{w} - 2\lambda \mathbf{w} = 0$, z čehož plyne

> $\mathbf{w}$ musí být vlastním vektorem $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$ odpovídajícím největšímu vlastnímu číslu.

3. při $s > 1$ postupujeme obdobně, promítáme do s-dimenzionální nadroviny

$$\mathbf{y}_i = \mathbf{x}_0 + \sum_{j=1}^{s} t_i^j \mathbf{w}^j$$

a zjistíme, že $\mathbf{w}^j$ jsou rovny **ortogonálním vlastním vektorům** $\mathbf{S}$ odpovídající s největším vlastním číslům.